RUHR

ECONOMIC PAPERS

Nolan Ritter
Colin Vance

The Phantom Menace of Omitted Variables

A Comment

RWI #282

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #282

Responsible Editor: Christoph M. Schmidt

ISSN 1864-4872 (online) – ISBN 978-3-86788-327-6

Ruhr Economic Papers #282

Nolan Ritter and Colin Vance

The Phantom Menace of Omitted Variables

A Comment

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

ISSN 1864-4872 (online)
ISBN 978-3-86788-327-6

Nolan Ritter and Colin Vance[1]

The Phantom Menace of Omitted Variables – A Comment

Abstract

This note demonstrates that in applied regression analysis, the variance of a coefficient of interest may decrease from the inclusion of a control variable, contrasting with Clarke's assertion (2005, 2009) that the variance can only increase or stay the same. Practitioners may thus be well-advised to include a relevant control variable on this basis alone, particularly when it is weakly correlated with the variable of interest.

JEL Classification: C12, C15, C18

Keywords: Control variables; variance; model specification

September 2011

1 *Nolan Ritter, RWI; Colin Vance, RWI and Jacobs University Bremen.* – *We thank Manuel Frondel for his discerning technical and editorial remarks. We are also indebted to Alfredo Payolo, Christoph M. Schmidt and Harald Tauchmann for helpful discussions during the drafting of this note.* – *All correspondence to Nolan Ritter, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI), Hohenzollernstr. 1-3, 45128 Essen, Germany. E-Mail: nolan.ritter@rwi-essen.de.*

Introduction

This note clarifies a point made in two papers appearing in *Conflict Management and Peace Science* by Clarke (2005, 2009) that explore the implications of omitted variable bias in regression analysis. Among the issues taken up by Clarke is whether the inclusion of relevant control variables decreases the variance of the coefficient of interest. He states that the quick answer is no, and elaborates:

Adding a variable [therefore] can never decrease the variance of the coefficient of interest; the variance can only increase or stay the same (Clarke 2005: 347; Clarke 2009: 48).

The aim of this note is to demonstrate why, in practice, the estimated variance of a coefficient of interest may well decrease from the inclusion of a control variable; whether this happens depends on the correlation of the included variable and the variable of interest.

Demonstration

Following Clarke's 2005 article, we assume that the true regression model is:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \quad (1)$$

Clarke then considers two misspecified models, Model 1 and 2, respectively:

Model 1: $Y_i = \beta_{11} + \beta_{21} X_{i2} + \epsilon_{i1}, \epsilon_{i1} \sim N(0, \sigma_1^2)$ (2)

Model 2: $Y_i = \beta_{12} + \beta_{22} X_{i2} + \beta_{32} X_{i3} + \epsilon_{i2}, \epsilon_{i2} \sim N(0, \sigma_2^2)$ (3)

and notes that the error variance of the OLS estimate $\hat{\beta}_{21}$ in Model 1 is given by:

$$Var(\hat{\beta}_{21}) = \frac{\sigma_1^2}{s_2} \quad (4)$$

where $s_2 = \sum_{i=1}^{n} (X_{i2} - \bar{X}_2)^2$. The error variance of the OLS estimate $\hat{\beta}_{22}$ in Model 2 is given by:

$$Var(\hat{\beta}_{22}) = \frac{\sigma_2^2}{s_2(1 - r_{23}^2)} \quad (5)$$

where r_{23} is the correlation coefficient of X_{i2} and X_{i3}. Implicitly assuming that σ_1^2 and σ_2^2 are equal, and recognizing that r_{23} falls somewhere along the 0 to 1 interval, Clarke concludes that the variance of $\hat{\beta}_{21}$ must necessarily be less than or equal to the variance of $\hat{\beta}_{22}$. This conclusion, however, is predicated upon a strong assumption that is rarely met in practice: σ_1^2 and σ_2^2 will only be equal in the

special case when X_{i3} adds no explanatory power to the regression in Model 2 ($\beta_{32} = 0$), which corresponds to the familiar result that the addition of irrelevant variables to a model unambiguously reduces the efficiency of the estimates.

But consider the difference in the variance estimate between Model 1 and Model 2 when X_{i3} is relevant. The formulae for the estimates of the residual variances are:

$$\hat{\sigma}_1^2 = \frac{\sum \hat{\epsilon}_{i1}^2}{n-2} \text{ and} \tag{6}$$

$$\hat{\sigma}_2^2 = \frac{\sum \hat{\epsilon}_{i2}^2}{n-3}. \tag{7}$$

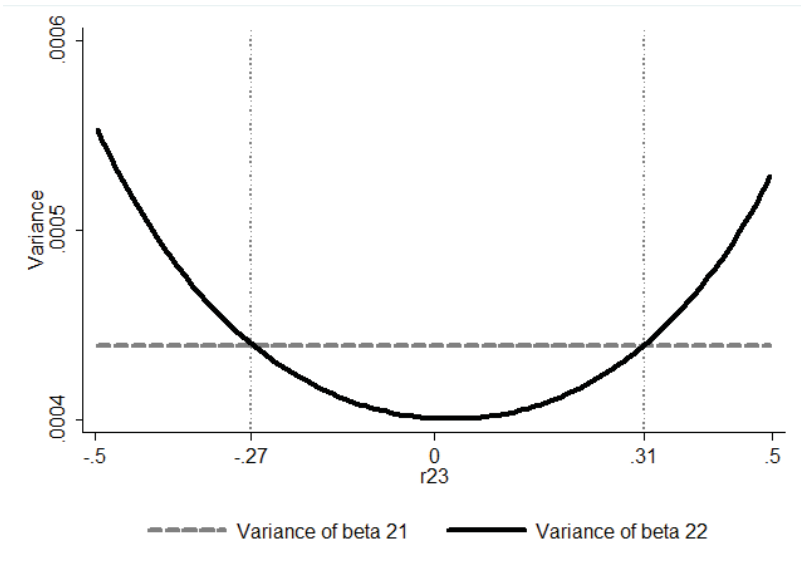
Further recognizing that $\epsilon_{i1} = \epsilon_i + \beta_3 X_{i3} + \beta_4 X_{i4}$ and that $\epsilon_{i2} = \epsilon_i + \beta_4 X_{i4}$, it becomes clear that there is no a priori reason for us to expect that σ_1^2 and σ_2^2 and, hence, the estimated residual variances, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, are equal. In fact, if X_3 is a relevant determinant of Y, it is quite plausible that σ_1^2 is greater than σ_2^2 because ϵ_1 contains part of X_3 (Wooldridge 2003: 101). Assuming this is the case, the question then turns to the magnitude of r_{23}. If r_{23} is relatively low, then a small drop in $\hat{\sigma}_2^2$ relative to $\hat{\sigma}_1^2$ would suffice to make the magnitude of the estimate for $Var(\hat{\beta}_{22})$ *less* than that of the estimate for $Var(\hat{\beta}_{21})$, contrasting with Clarke's theoretical assertion.

Does this occur with any frequency in applied empirical work? Absolutely, which is one reason Angrist and Pischke (2009: 24) give for including control variables: to reduce the residual variance, thereby lowering the estimated variance (and hence the estimated standard errors) of the regression estimates. Using observed data on food demand, for example, Maddala (2001: 161) presents a model that demonstrates how the variance may decrease with the inclusion of a control variable. In footnote 4 of his 2009 paper, Clarke also acknowledges that the estimated variance is biased when the restriction in Model 1 is false. But he maintains that this bias is rarely large enough to affect his findings.

To illustrate a counterexample, we implement a simple Monte Carlo experiment for which annotated code, written using Stata, is included in the appendix. Begin by assuming that the true model given by equation (1) is the data generation process. Setting the population to 10,000 observations, we randomly draw from a uniform distribution to generate values for X_2 and do likewise for X_4. Values for X_3 are generated by drawing from a uniform distribution and adding to this the product of a scalar *pi* and X_2, thereby allowing us to adjust the degree of correlation between $X_2$and X_3 depending on the magnitude of *pi*. The error term ϵ_i is drawn from a normal distribution with a mean of zero and variance of one. Inserting these variables into the true model, setting $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$, and selecting a value for *pi*, we generate 10,000 values of Y. The simulation then proceeds by drawing a 75% sample of the generated data and estimating Models 1 and 2. The process is repeated 1,000 times, yielding a distribution of estimates for $\hat{\beta}_{21}$ and $\hat{\beta}_{22}$, from which estimates of their respective variances can be calculated.

The graph below plots estimates of $Var(\hat{\beta}_{21})$ and $Var(\hat{\beta}_{22})$ from distinct runs of the simulation, with $Var(\hat{\beta}_{22})$ varying depending on the value of *pi* and hence the correlation between X_2 and X_3. In this example, when the correlation is lower than -0.27, the inclusion of variable X_3 in Model 2 makes the estimated variance of $\hat{\beta}_{22}$ higher than that of $\hat{\beta}_{21}$. When the correlation falls between -0.27 and 0.31, with the corresponding value of *pi* falling between -0.28 and 0.33, the opposite holds: $Var(\hat{\beta}_{22}) < Var(\hat{\beta}_{21})$. Finally, when the correlation is above 0.31, we again have the case in which $Var(\hat{\beta}_{22}) > Var(\hat{\beta}_{21})$. Thus, we see that for some 29% of the range in correlations between 1 and -1, the estimated variance of $\hat{\beta}_{22}$ is reduced from the inclusion of the control variable.[2]

Figure 1: Monte Carlo Simulations



Conclusion

One of the key points made in Clarke's highly insightful analysis is that there is a disconnect between textbook discussions of omitted variable bias and the real world confronted by practitioners. As he persuasively argues, the standard formula for omitted variable bias is of little use when the correct specification of the model is unknown; without this knowledge, the analyst is unable to determine whether adding a control variable or set of controls makes the bias on the coefficient of interest better or worse.

[2] Clarke (2005) additionally argues that whenever the bias on $\hat{\beta}_{21}$is less than the bias on $\hat{\beta}_{22}$, the mean squared error (MSE) of $\hat{\beta}_{21}$must be less than the MSE of $\hat{\beta}_{22}$. This assertion can also be contradicted with Monte Carlo simulation: If $Var(\hat{\beta}_{21})$ is greater than $Var(\hat{\beta}_{22})$, then the MSE of $\hat{\beta}_{21}$, may be greater than the MSE of $\hat{\beta}_{22}$even when the bias on the former is greater than the bias on the latter. Demonstration code is available upon request.

In this note, we suggest that the analyst likewise runs the risk of falling into the void between theory and practice when assessing the effect of the inclusion of a control variable on the variance of the estimate of the coefficient of interest. Clarke argues that this variance can only increase or stay the same. While this result is always true for the special case in which the coefficient on the control variable is zero, our analysis shows that the estimated variance may well decrease in the more realistic case when the control variable explains variation in Y, which has relevance for the specification of the model. Specifically, if controls are available that are weakly correlated or uncorrelated with the variable of interest, but are relevant determinants of the dependent variable, a circumstance that is not uncommon in natural experiments (e.g. Galiani and Gertler 2005; Little, Long and Lin 2009), then their inclusion can serve to increase the precision of the coefficient estimates without imparting bias.

References

Angrist, Joshua and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Clarke, Kevin A. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22: 341-352.

Clarke, Kevin A. 2009. Return of the phantom menace: Omitted variable bias in political research. *Conflict Management and Peace Science*, 26: 46-66.

Galiani, Sebastian and Paul Gertler. 2005. Water for life: The impact of the privatization of water services on child mortality. *Journal of Political Economy* 113: 83-120.

Little , Roderick J., Qi Long, and Xihong Lin. 2009. A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics* 65 (June): 640–649.

Maddala, G.S. 2001. *Introduction to Econometrics*. Third Edition. Chichester: John Wiley & Sons, Ltd.

Wooldridge, Jeffrey M. 2003. *Introductory econometrics: A modern approach*. Second Edition. Mason: Thomson South-Western.

Appendix

The following presents annotated Stata code, Version 10.0, for implementing the Monte Carlo simulation. Users can change the value of pi, currently set at 0.325, to explore the implications of different degrees of correlation (r23) between x2 and x3.

```
*** PROGRAMMING OF MONTE-CARLO-ROUTINE STARTS
capture program drop monte  /* drops any existing components of program "monte" to assure a clean start */
program monte, rclass       /* program monte definition begins */
version 10.0                /* command interpreter set to STATA version 10.0 */
capture drop uniform
gen uniform = uniform()             /* the variable "uniform" is drawn from a uniform distribution with values between 0 and 1 */
regress y x2 if uniform > 0.25      /* regresses y on x2 on a subsample where the variable "uniform" is greater than 0.25 */
return scalar beta_21 = _b[x2]      /* stores the estimated coefficient of beta_21 */
regress y x2 x3 if uniform > 0.25   /* regresses y on x2 and x3 on the same sample as above */
return scalar beta_22 = _b[x2]      /* stores the estimated coefficient of beta_22 */
corr x2 x3                          /* calculates the empirical correlation between x2 and x3 */
return scalar correlation = r(rho)  /* stores the correlation */
regress y x2 x3 x4 if uniform > 0.25  /* regresses y on x2, x3 and x4 on the same sample as above */
return scalar beta_2 = _b[x2]       /* stores the estimated coefficient of beta_2 */
end                                 /* definition of program monte ends */
*** CREATE DATASET
clear
set seed        1111        /* random number generator set to seed 1111 to make sure the dataset is always the same */
set obs         10000       /* sets the number of observations to 10000  */
capture drop beta_1 x2 x3 x4 y
generate beta_1 = 1                     /* the intercept takes on the value 1 */
generate double x2 = uniform()          /* x2 is drawn from a uniform distribution with values between 0 and 1 */
generate double x3 = uniform() + 0.325 * x2  /* pi is set at 0.325. */
generate double x4 = uniform()          /* x4 is drawn from a uniform distribution with values between 0 and 1 */
generate double y = beta_1 + x2 + x3 + x4 + rnormal(0,1) /* build y from generated components assuming slope coefficients of 1 plus a
normally distributed error term with expected value 0 and variance 1 */
```

```
*** SIMULATION BEGINS
simulate correlation = r(correlation) beta_2 = r(beta_2) beta_21 = r(beta_21) beta_22 = r(beta_22), reps(1000): monte /* runs monte-carlo simulation with 1000 repetitions */
*** LABEL VARIABLES
label var beta_21 "coefficient estimates for beta_21"
label var beta_22 "coefficient estimates for beta_22"
label var beta_2  "coefficient estimates for beta_2 from full model"

*** GET VARIANCE ESTIMATES
sum beta_21, detail                /*Compare presented variance estimate with that of beta_22.*/
sum beta_22, detail                /*Compare presented variance estimate with that of beta_21.*/
display correlation


/* To generate the data to make the graph presented in the paper, repeat the Monte-Carlo simulation for various values of pi and store the results for the correlation of x2 and x3 (r23) and the variances of the betas along with pi in a separate data set. Additional code to automate this process is available from the authors upon request. */
```